Standard oder Fehler? Einige Eigenschaften von Schätzverfahren bei komplexen Stichprobenplänen und aktuelle Lösungsansätze

Lipsmeier, Gero

Standard oder Fehler? Einige Eigenschaften von Schätzverfahren bei komplexen Stichprobenplänen und aktuelle Lösungsansätze[1]

von Gero Lipsmeier[2]

Zusammenfassung

Datenanalysen in den Sozialwissenschaften beruhen fast ausschließlich auf Stichproben. Dabei sind einfache, uneingeschränkte Zufallsstichproben ausgesprochen selten. In der Regel basieren die ausgewerteten Daten auf komplexen Stichprobenplänen mit mehrstufigen Auswahlverfahren sowie Schichtung und Klumpung. Während die Verwendung von Gewichtungsfaktoren mittlerweile weit verbreitet ist, bleibt die Stichprobenanlage meistens unberücksichtigt. Dieser Beitrag zeigt, daß dieses Standardvorgehen durchaus ein Fehler sein kann. Neben einem knappen Überblick über die einschlägige statistische Theorie wird an mehreren Beispielen demonstriert, daß Teststatistiken oft fälschlicherweise signifikante Ergebnisse suggerieren, wenn das Design der Stichprobe nicht berücksichtigt wird. Diesen Beispielen liegt ein aktueller Datensatz nach dem verbreiteten ADM-Auswahlverfahren zugrunde, und es wird gezeigt, wie sich die Stichprobenstruktur dieser Daten mit aktueller Statistiksoftware berücksichtigen läßt.

Abstract

Data analysis in the social sciences is almost uniformly based on samples and simple, unrestricted random samples are extremely rare. Generally the analysed data has its origin from complex sampling designs, involving stratification and clustering. While the use of weighting factors is now quite common, the sampling design itself is frequently disregarded. This article shows that this standard procedure may well be a mistake. Besides giving a brief overview on the relevant statistical theory, a couple of examples are used to demonstrate that teststatistics often falsely suggest significant results if the sampling design is not accounted for. For these examples a recent data set from the widely used ADM-sampling procedure is used and it is shown, how this design can be accounted for with recent statistical software.

1 **Für wertvolle Hinweise und Anregungen danke ich** *Siegfried Gabler* **und** *Karin Kurz***.**
2 *Gero Lipsmeier***, Universität Bielefeld, Fakultät für Soziologie, Postfach 100 131, 33501 Bielefeld.**

1. Einleitung

In diesem Beitrag möchte ich zum einen zeigen, daß es wichtig ist, die Stichprobenstruktur nicht nur in Form von Gewichtungsfaktoren zu berücksichtigen. Es wird deutlich werden, daß inferenzstatistische Schlüsse ansonsten häufig zu fälschlicherweise signifikanten Ergebnissen führen, da die geschätzten Standardfehler zu klein sind. Zum anderen möchte ich demonstrieren, daß es unter bestimmten Voraussetzungen mit moderner Software technisch leicht möglich ist, bessere Schätzer für die Standardfehler vieler Statistiken zu erlangen. Insofern ist dieser Beitrag auch als Plädoyer für die häufigere Verwendung von entsprechenden Verfahren gemeint.

Komplexe Stichproben mit (disproportionaler oder proportionaler) *Schichtung* und/oder *Klumpung* sowie mehreren *Auswahlstufen* sind in der Praxis der empirischen Sozialforschung weitaus häufiger anzutreffen als einfache Zufallsauswahlen. Nahezu jedes Lehrbuch der empirischen Sozialforschung nennt hierfür eine Vielzahl guter Gründe. Als Vorteile von geschichteten Stichproben gegenüber einfachen Zufallsstichproben führen z.B. ***Schnell et al.*** (1992: 295) an, daß a) die Schätzung mit geschichteten Stichproben *genauer sein kann*, wenn sich die Schichten in der Grundgesamtheit bei der Streuung des interessierenden Merkmales unterscheiden, b) geschichtete Stichproben kostengünstiger sein können, wenn es Unterschiede zwischen den Schichten in Hinsicht auf die Kosten der Erhebung gibt und c) unabhängige Schätzungen für jede Schicht erfolgen können, wenn die Schichten selbst von Interesse sind. Letzteres Argument ist häufig Motivation für eine disproportionale Schichtung nach bestimmten interessierenden Merkmalen. Geläufige Beispiele hierfür sind die Überrepräsentation von Ausländern im Sozioökonomischen Panel oder die getrennte Erhebung von Ost- und Westdeutschland mit überproportionaler Berücksichtigung der neuen Bundesländer, wie z.B. in allen ALLBUS-Umfragen seit der deutschen Vereinigung. Mit disproportionaler Schichtung geht einher, daß die Auswahlwahrscheinlichkeiten der Grundgesamtheitselemente - im Gegensatz zur einfachen Zufallsstichprobe - nicht identisch sind. Demzufolge erfordert die Bestimmung von unverzerrten Schätzern für Grundgesamtheitsparameter (wie z.B. das arithmetische Mittel oder Regressionskoeffizienten) die Verwendung von *Gewichten.*

Ebenso zwingt das Fehlen eines zentralen Einwohnerregisters für die Bundesrepublik die Forscher und Forscherinnen bei Erhebungen der allgemeinen Bevölkerung in der Regel zur Verwendung von mehrstufigen Auswahlverfahren mit einer geographischen Klumpung. Von Klumpenstichproben ist allgemein bekannt, daß die Schätzung von Parametern der Grundgesamtheit in der Regel *ungenauer* ist als bei einfachen Zufallsstichproben. Das gilt insbesondere dann, wenn sich die Elemente innerhalb der Klumpen (z.B. Wahlbezirke) ähn-licher sind als dieses bei einer einfachen Zufallsstichprobe zu erwarten wäre. „Klumpenstichproben führt man aus praktischen Gründen durch, wenn eine einfache Zufallsauswahl nicht möglich oder zu aufwendig ist. Dabei nimmt man einen Verlust an Präzision in Kauf.„ (***Diekmann*** 1995: 338).

Häufig sind mehrstufige Auswahlverfahren so konstruiert, daß innerhalb der einzelnen Schichten *Haushalte* mit der jeweils gleichen Auswahlwahrscheinlichkeit gezogen werden. Wird nur eine *Person* im Haushalt befragt, so wird diese durch eine weitere Stufe des Auswahlverfahrens (z.B. durch den Einsatz von Schwedenschlüsseln) ausgewählt. Die Auswahlwahrscheinlichkeit von Personen ist somit auch von der Anzahl der Personen im Haushalt abhängig, die zur angestrebten Grundgesamtheit zu rechnen sind. Somit ist bei Analysen auf Personenebene häufig ebenfalls die Verwendung von Gewichtungsfaktoren erforderlich. Weiterhin werden Gewichtungsfaktoren oft verwendet, um die Verteilung einiger zentraler Merkmale in der (gewichteten) Stichprobe nachträglich an bekannte Verteilungen in der Grundgesamtheit anzugleichen (sogenanntes Redressement).

All dies ist für Sozialforscher, die mit Umfragedaten arbeiten, nichts Neues, und die Verwendung von Gewichtungsfaktoren zum Ausgleich unterschiedlicher Auswahlwahrscheinlichkeiten und/oder als Redressement ist seit längerem als Standard etabliert. Jede halbwegs aktuelle Statistiksoftware erlaubt dementsprechend die Verwendung von Gewichtungsvariablen bei fast allen Datenanalysen. Die Konsequenzen des *Stichprobenplans* für die Genauigkeit der Schätzung werden jedoch weitgehend ignoriert. In fast allen Aufsätzen und Monographien, in denen Datenanalysen auf der Basis von komplexen Stichproben berichtet werden, wird die Schätzung von Standardfehlern und daraus abgeleiteten Teststatistiken umstandslos der eingesetzten Statistiksoftware überlassen. Kaum jemand legt Rechenschaft darüber ab, daß damit in aller Regel die *Formeln für einfache Zufallsstichproben* zum Einsatz kommen. Während durch den Einsatz von Gewichten unverzerrte Punktschätzer bestimmt werden können, werden die Standardfehler in der Regel so bestimmt, als läge den Schätzungen eine einfache (wenn auch gewichtete) Zufallsstichprobe zugrunde.

Dieses ‚Standardvorgehen' soll an einem einfachen Beispiel demonstriert werden: Angenommen uns interessiere das durchschnittliche persönliche Nettoeinkommen von deutschen, in Privathaushalten der Bundesrepublik Deutschland lebenden Personen, die Ende 1996 das 18. Lebensjahr vollendet hatten[3]. Nehmen wir weiter an, daß wir eine *einfache Zufallsstichprobe* im Umfang von n = 2.600 Personen aus der so definierten Grundgesamtheit zu ihrem persönlichen Nettoeinkommen befragt hätten. In diesem Fall ist

das arithmetische Mittel der Einkommensangaben x_i (i = 1,2,...,n) in der Stichprobe

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1.861{,}32 \text{ DM}$$

3 So ist die Grundgesamtheit für den Sozialwissenschaftenbus III/96 definiert (vgl. GFM-GETAS/WBA 1997: 4). Für alle Beispiele dieses Beitrages verwende ich diese Datenquelle. Die Einschaltung in den Sozialwissenschaften-Bus wurde durch einen Studienpreis, den GFM-GETAS/WBA anläßlich seines 50jährigen Jubiläums gestiftet hat, und aus Mitteln des Ministeriums für Wissenschaft und Forschung des Landes Nordrhein-Westfalen finanziert. Diese Förderung wurde vom Autor gemeinsam mit ***Hans-Jürgen Andreß*** eingeworben. Weitere Erläuterungen zum zugrundeliegenden Auswahlverfahren finden sich im folgenden Abschnitt.

ein unverzerrter Schätzer für das Durchschnittseinkommen μ in der Grundgesamtheit[4]. Ebenso ist die Varianz der Einkommensangaben in der Stichprobe

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = 2.304.867,48$$

ein geeigneter Schätzer für die Merkmalsvarianz σ^2 der Einkommen in der Grundgesamtheit. Aus dieser wiederum gewinnen wir (unter Vernachlässigung des Korrekturfaktors für endliche Grundgesamtheiten) einen Schätzer für den Standardfehler des arithmetischen Mittels:

$$\hat{\sigma}_{\bar{X}} = \sqrt{s^2 \frac{1}{n}} = 29,44$$

Da es sich (hypothetisch) um eine einfache Zufallsstichprobe handelt und der Stichprobenumfang hinreichend groß ist, kann man z.B. ein 95% Konfidenzintervall für das Durchschnittseinkommen in der Grundgesamtheit bestimmen:

$$\text{untere Grenze}: \bar{x} - 1,96\,\hat{\sigma}_{\bar{X}} = 1.861,32 - 1,96 * 29,44 = 1.803,60$$
$$\text{obere Grenze}: \bar{x} + 1,96\,\hat{\sigma}_{\bar{X}} = 1.861,32 + 1,96 * 29,44 = 1.919,04$$

Einfache Zufallsstichproben sind in der Praxis der empirischen Sozialforschung jedoch ausgesprochen selten. Deshalb wollen wir dieses einfache Beispiel um die Berücksichtigung von unterschiedlichen Auswahlwahrscheinlichkeiten erweitern. Unterstellen wir zu diesem Zweck, daß die betrachtete Stichprobe aus zwei getrennten Stichproben für Ost- und Westdeutschland besteht. Dabei seien die ostdeutschen Befragten mit einer höheren Auswahlwahrscheinlichkeit in die Stichprobe aufgenommen als die Westdeutschen. Es handelt sich also um eine disproportional nach der Region geschichtete Stichprobe. Offensichtlich ist das einfache arithmetische Mittel in diesem Fall *kein* unverzerrter Schätzer für das Durchschnittseinkommen in Gesamtdeutschland. Durch die überproportionale Berücksichtigung von (ostdeutschen) Personen mit durchschnittlich niedrigeren Einkommen würde das Durchschnittseinkommen in der Grundgesamtheit unterschätzt. Wie bereits angedeutet, läßt sich dieses Problem lösen, indem man nicht ein einfaches arithmetisches Mittel berechnet, sondern ein *gewichtetes Mittel*. Die Gewichte werden dabei so konstruiert, daß sie die

4 Bei diesen Überlegungen lasse ich den komplexen Bereich von (nichtzufälligen) Antwortausfällen und Antwortverweigerungen unberücksichtigt. Thema dieses Beitrages ist ausschließlich der Einfluß des Stichprobendesigns. Weiterhin wurde für die Berechnung des hier ausgewiesenen Durchschnittseinkommens ein (in der Formel nicht dargestelltes) Gewicht zur Anpassung an Analysen auf Personenebene, aber *ohne* Korrektur für die Überrepräsentation Ostdeutscher verwendet, um die Vergleichbarkeit der numerischen Ergebnisse mit den folgenden Berechnungen sicherzustellen.

unterschiedlichen Auswahlwahrscheinlichkeiten ausgleichen. Definieren wir w_i als das zu jeder Beobachtung i gehörende Gewicht[5], so erhalten wir:

$$\bar{x}_w = \frac{1}{W} \sum_{i=1}^{n} w_i \, x_i = 1.892{,}85 \text{ DM}$$

$$\text{mit} \quad W = \sum_{i=1}^{n} w_i = 2.660$$

Analog zum Vorgehen bei einer einfachen Zufallsstichprobe läßt sich auch die Merkmalsvarianz unter Berücksichtigung von Gewichten bestimmen:

$$s_w^2 = \frac{1}{W-1} \sum_{i=1}^{n} w_i \, (\bar{x}_w - x_i)^2 = 2.535.922{,}48$$

Standardmäßig wird hieraus analog zum oben wiedergegebenen Vorgehen bei einfachen (ungewichteten) Zufallsstichproben der Standardfehler des arithmetischen Mittels abgeleitet:

$$\hat{\sigma}_{\bar{x}_w} = \sqrt{s_w^2 \frac{1}{W}} = 30{,}88$$

Genau nach dieser Formel wird der Standardfehler von den gebräuchlichen Statistikpaketen berechnet, wenn Gewichtungsvariablen spezifiziert werden[6]. Offensichtlich finden Informa

5 Die vom Umfrageinstitut bereitgestellte Gewichtungsvariable wurde für dieses Beispiel so reskaliert, daß die Summe der Gewichte W dem Stichprobenumfang n entspricht. Weiterhin beinhalten die bereitgestellten Gewichte auch Anpassungen (Redressment) an Randverteilungen in der Grundgesamtheit und eine Designgewichtung für Analysen auf Personenebene. Hier soll jedoch lediglich die ebenfalls erreichte Korrektur für die Überrepräsentation Ostdeutscher betrachtet werden.

6 STATA (5.0), SPSS (7.5.2) und SAS (6.12) unterscheiden sich jedoch in den Standardvorgaben bei unterschiedlichen Skalierungen der Gewichtungsvariablen. Die Verwendung von sogenannten ‚Analytical Weights' [aweights] führt bei STATA dazu, daß die Gewichtungsvariable intern so skaliert wird, daß die Summe der Gewichte dem Stichprobenumfang entspricht. Dadurch sind sowohl geschätzte Merkmalsvarianz (bzw. Standard-abweichung) als auch Standardfehler von der tatsächlichen Skalierung unabhängig. SPSS verwendet die oben angegebenen Berchnungsformeln *ohne* vorherige Anpassung der Skalierung. Wie man an den Formeln leicht erkennt, ist somit der geschätzte Standardfehler (nicht hingegen die Merkmalsvarianz) von der Skalierung ab-hängig. Demgegenüber verwendet SAS als Standardvorgabe bei der Berechnung der Merkmalsvarianz nicht 1/W-1 sondern 1/n-1 (SAS Institute Inc**.** 1990: 375). Damit ist die geschätzte Merkmalsvarianz skalierungsabhängig. Bei der Berechnung des Standardfehlers wird jedoch wieder durch die Summe der Gewichte W geteilt, so daß der Standardfehler *nicht* von der Skalierung der Gewichtungsvariablen abhängt. SAS ermöglicht die Beeinflussung der Berechnungsweise über Optionen.

tionen über die Schichtung der Stichprobe keine explizite Berücksichtigung bei der Berechnung.

Sollte man aber nicht erwarten, daß der Schätzer für den Standardfehler mögliche Genauigkeitsgewinne durch eine gelungene Schichtung berücksichtigt? Welchen Einfluß hat das Gewichtungsverfahren selbst auf die Präzision von Schätzern, und ist die einfache Übertragung der Berechnungsformeln für einfache Zufallsstichproben auf gewichtete Schätzer durch die statistische Theorie abgesichert? Ebenso müßte man von einer Berechnungsanweisung für das Maß der Genauigkeit von Schätzergebnissen - nichts anderes sind Formeln für Standardfehler schließlich - erwarten, daß es mögliche Genauigkeitsverluste von Schätzungen bei Klumpenstichproben durch größere Standardfehler reflektiert. Auch Klumpungseffekte finden in der obigen Formel keine Berücksichtigung und werden bei dieser Berechnungsweise dementsprechend ignoriert. Wie ist schließlich die Präzision von Schätzern zu quantifizieren, wenn Schichtung, Klumpung und unterschiedliche Auswahlwahrscheinlichkeiten - wie in mehrstufigen Designs üblich - simultan vorkommen? Als Antwort auf diese Fragen kennt die statistische Theorie seit langem verschiedene Verfahren zur Bestimmung der Varianz vieler gebräuchlicher Schätzer unter Berücksichtigung der Stichprobenstruktur. Leider waren derartige Prozeduren bis vor kurzem jedoch nicht in der in den Sozialwissenschaften verbreiteten Standardsoftware wie SPSS und SAS implementiert, sondern allenfalls als zusätzlich zu beschaffende Spezialsoftware für begrenzte Anwendungsgebiete verfügbar. Hierin ist sicher ein wichtiger Grund für die geringe Aufmerksamkeit zu sehen, die solche Verfahren bislang gefunden haben. Diese Situation beginnt sich jedoch allmählich zu ändern. Seit Version 5 der Statistiksoftware Stata stehen entsprechende Prozeduren in einem leistungsfähigen Statistikpaket standardmäßig zur Verfügung. Weiterhin vertreibt SPSS mittlerweile exklusiv die Spezialsoftware WesVar, die sich in SPSS für Windows integriert und somit die dort bereitgestellten Funktionen zur Datenmanipulation nutzen kann. Mit Sudaan (http://www.rti.org/patents/sudaan/sudaan.html) steht sowohl eine eigenständige Software zu Verfügung als auch eine Variante, die auf eine Zusammenarbeit mit SAS eingerichtet ist

Im folgenden Abschnitt möchte ich zunächst das Auswahlverfahren des Sozialwissenschaftenbusses III/1996 als ein typisches und oft eingesetztes Beispiel für ein komplexes, mehrstufiges Auswahlverfahren vorstellen. Alle empirischen Ergebnisse dieses Beitrages basieren - wie das einleitende Beispiel - auf diesen Daten. Daran anschließend (3. Abschnitt) soll unter etwas stärkerem Rückgriff auf die statistische Theorie gezeigt werden, wodurch die Schätzung von Standardfehlern bei komplexen Stichproben gegenüber einfachen Zufallsstichproben verkompliziert wird. Ebenso sollen in diesem Abschnitt einige gängige Lösungsstrategien aufgezeigt werden. Im 4. Abschnitt soll schließlich an zwei konkreten Beispielen demonstriert werden, daß es tatsächlich einen erheblichen Unterschied für inferenzstatistische Schlüsse machen kann, ob die Standardfehler nach der oben dargestellten

Standardmethode oder mit Verfahren berechnet werden, die das Stichprobendesign berücksichtigen.

2 Ein Beispiel für eine komplexe Stichprobe: Das ADM-Stichprobendesign des Sozialwissenschaftenbusses

Das Auswahlverfahren für die Stichprobe des von ZUMA-Mannheim e.V. und GFM-GETAS/WBA gemeinsam mehrmals pro Jahr durchgeführten Sozialwissenschaftenbusses ist ein recht typisches Beispiel für einen komplexen Stichprobenplan. Ähnliche Designs wurden und werden bei Bevölkerungsumfragen in der Bundesrepublik häufig eingesetzt (z.B. bei den ALLBUS-Umfragen bis einschließlich 1992). Für den Sozialwissenschaftenbus III/96 wurde eine Stichprobe von insgesamt n = 3.170 Personen aus der Grundgesamtheit von deutschen, in Privathaushalten der Bundesrepublik lebenden, Personen im Alter von 18 Jahren an realisiert. Von diesen lebten zum Befragungszeitpunkt n = 1.989 in den westlichen Bundesländern und n = 1.181 in den östlichen Bundesländern. Die Bruttoausgangsstichprobe(Anzahl der eingesetzten Adressen)betrug n =3.360/West sowie n= 1.680/-Ost. Vergleicht man die realisierten Stichprobenumfänge mit der Anzahl von erwachsenen Personen in den alten bzw. neuen Ländern, so erkennt man leicht die (beabsichtigte) überproportionale Berücksichtigung von Personen aus den neuen Bundesländern. Nach der Übersicht des statistischen Bundesamtes (***Statistisches Bundesamt*** 1998: 37) lebten am 31.12.1996 insgesamt 54.755.936 Personen über 18 Jahren in den alten Bundesländern und 11.335.075 Erwachsene in den neuen Bundesländern[7]. Somit haben Personen aus den neuen Bundesländern eine ca. 2,5mal so hohe Chance in die Stichprobe zu gelangen. Die Stichprobe ist demnach *disproportional nach dem Merkmal alte versus neue Bundesländer geschichtet*. In beiden Schichten wurden unabhängige Stichproben nach dem im folgenden beschriebenen mehrstufigen Auswahlverfahren gezogen.

Auf der ersten Auswahlstufe wurden auf der Basis von ADM-Stichprobennetzen (vgl. ***Kirschner*** 1984a, ***Arbeitsgemeinschaft ADM-Stichproben*** und ***Bureau Wendt*** 1994) 630 (420/West, 210/Ost) Wahlbezirke durch eine systematische Zufallsauswahl gezogen. Das Ziehungsverfahren war dabei durch eine geeignete strukturierte Anordnung der Wahlbezirke (Sampling Points) so gestaltet,daß die Auswahlwahrscheinlichkeit jedes einzelnen Sampling Points proportional zur durch ein sogenanntes Bedeutungsgewicht quantifizierten geschätzten Anzahl von Privathaushalten war. Wird deshalb in den anschließenden Auswahlschritten in jedem der gezogenen Sampling Points die gleiche Anzahl von Haushalten ausgewählt, so garantiert dieses Verfahren (innerhalb der jeweiligen Schicht) identi

7 Die Verwendung der Bevölkerungszahlen als Vergleichsgröße überschätzt den Umfang der Grundgesamtheit etwas, da Ausländer und Personen, die nicht in Privathaushalten leben, nicht zur definierten Grundgesamtheit zählen. Ignoriert man diese Ungenauigkeit, so betrug der Auswahlsatz für Westdeutschland ca. 1.989/54.755.936 = 0,000036 und in den neuen Bundesländern ca. 1.181/11.335.075 = 0,000104.

sche Auswahlwahrscheinlichkeiten für alle *Haushalte*. Derartige Auswahlverfahren bezeichnet man üblicherweise als PPS-Designs (probability proportional to size). Korrekterweise müßte man dieses Verfahren jedoch als PPES-Design (probability proportional to *estimated* size) bezeichnen, da die Anzahl der Privathaushalte in den einzelnen Wahlbezirken nicht fehlerfrei feststellbar ist, sondern geschätzt werden muß. Diese Unterscheidung wird uns weiter unten (Abschnitt 3) noch beschäftigen.

Auf der zweiten Auswahlstufe wurden in jedem der ausgewählten Sampling Points Haushalte mit einem Random-Route-Verfahren ausgewählt. Ausgehend von einer zufällig ausgewählten Startadresse innerhalb der Sampling Points waren die Interviewer gehalten, nach einer Begehungsanweisung insgesamt 23 Adressen von Privathaushalten im räumlichen Abstand von drei Haushalten aufzulisten. Die aufgelisteten Adressen liegen somit in unmittelbarer Nachbarschaft zueinander. Aus den aufgelisteten Adressen wurden auf dieser Auswahlstufe schließlich jeweils 8 Haushalte (per systematischer Zufallsauswahl) ausgewählt, in denen Interviews durchgeführt werden sollten.

Als letzte Stufe des Auswahlprozesses wurde schließlich die in den jeweiligen Haushalten zu befragende Personen mit einem Schwedenschlüssel-Verfahren ausgewählt. Da die vorangegangenen Auswahlstufen (innerhalb der Schichten) für *Haushalte* die gleiche Auswahlwahrscheinlichkeit sicherstellen, ist die Auswahlwahrscheinlichkeit von Personen aus Haushalten mit mehreren zur Grundgesamtheit gehörenden Personen abhängig von der Anzahl dieser Personen.

Aus stichprobentheoretischer Sicht läßt sich diese Stichprobe somit als nach dem Schichtungsmerkmal Ost-West disproportional geschichtete Stichprobe von 630 Wahlbezirken als primäre Stichprobeneinheiten (primary sampling units, PSUs) beschreiben, in denen jeweils ein zweistufiges Zufallsauswahlverfahren zur Bestimmung der insgesamt 5.040 zu befragenden Personen (Bruttoadressenansatz) durchgeführt wurde. ***Kirschner*** (1984b) beschreibt dieses Auswahlverfahren als eine Abfolge mehrerer Zufallsprozesse, die es bei der Bestimmung von Fehlervarianzen auf der Basis dieser Stichprobe zu berücksichtigen gilt. In dem zitierten Beitrag ist der Aspekt der Schichtung allerdings noch nicht berücksichtigt, da das ADM-Auswahlverfahren für die alte Bundesrepublik keine disproportionale Schichtung aufwies.[8]

8 Streng genommen war (ist) das ADM-Stichprobendesign duch die strukturierte Anordnung der primären Ziehungseinheiten auch vor der deutschen Vereinigung ein (proportional) geschichtetes Design. Weiterhin kann man einwenden, daß die Zufälligkeit der Auswahl auf der zweiten Auswahlstufe durch die vorherige Markierung von zu befragenden Haushalten in den Adreßauflistungsbögen beeinträchtigt wird. Um die Argumentation nicht zusätzlich zu verkomplizieren, lasse ich beide Aspekte hier unberücksichtigt.

3 Ein kurzer Einblick in die statistische Theorie zum Zusammenhang von Stichprobendesigns und Eigenschaften von Schätzern

Die oben (Abschnitt 1) dargestellten Formeln für Standardfehler von Schätzern für Grundgesamtheitsparameter sind *nur* für einfache (ungewichtete) Zufallsstichproben (mit Zurücklegen) korrekt. Sowohl Schichtung als auch Klumpung beeinflussen die Varianz von Schätzern in unterschiedlicher Weise. Wir benötigen deshalb Varianzschätzer, die diese simultanen (und möglicherweise gegenläufigen) Effekte berücksichtigen. Im folgenden wollen wir uns deshalb zunächst mit den durch Schichtung und Klumpung zu erwartenden Effekten auf die Größe von Standardfehlern beschäftigen. Dabei beschränke ich mich der Anschaulichkeit halber auf die Schätzung des Standardfehlers für das arithmetische Mittel. Die grundlegenden Prinzipien sind jedoch auf viele andere Schätzer übertragbar. Daran anschließend wird zu diskutieren sein, inwieweit die Tatsache, daß bei komplexen Stichprobenverfahren der theoretisch zu erwartende Stichprobenumfang als das Ergebnis eines Zufallsprozesses zu betrachten ist, die Schätzung von Standardfehlern weiter verkompliziert. Zum Abschluß dieses Abschnitts soll kurz auf besondere Eigenschaften von Maximum-Likelihood-Schätzern bei komplexen Stichproben eingegangen werden.

(Disproportionale) Schichtung

Die oben dargestellte gewichtete Berechnungsweise für einen unverzerrten Schätzer des arithmetischen Mittels μ läßt sich auch als gewichtete Summe der schichtspezifischen Mittelwerte ausdrücken. Bezeichnet man den Umfang der Grundgesamtheit mit N und den Umfang jeder Schicht mit N_h (mit h = 1,2,...,H; H = Anzahl der Schichten), so beträgt der Anteil der Schicht an der Grundgesamtheit $N_h / N = W_h$. Die Summe dieser Anteile über alle Schichten beträgt somit $\Sigma W_h = 1$. Die mit W_h gewichtete Summe aus dem für jede Schicht getrennt berechneten arithmetischen Mittel $\bar{x}_h$ über alle H-Schichten ist dann der gesuchte unverzerrte Schätzer für das arithmetische Mittel in der Grundgesamtheit:

$$\hat{\mu} = \bar{x}_{gesch} = \sum_{h=1}^{H} W_h \bar{x}_h$$

Da das geschichtete Auswahlverfahren eine unabhängige Auswahl in den einzelnen Schichten liefert, kann die Varianz dieses Schätzers wie folgt geschätzt werden (vgl. ***Kish*** 1965: 78; ***Stenger*** 1986: 116

$$\hat{\sigma}^2_{\bar{x}_{gesch}} = \text{var}\left(\sum_{h=1}^{H} W_h \bar{x}_h \right) = \sum_{h=1}^{H} W_h^2 \ \text{var}\,(\bar{x}_h)$$

An dieser Berechnungsweise sind zwei Dinge bemerkenswert:Zum einen geht die Streuung der Merkmale in den einzelnen Schichten getrennt in die Berechnung der gesamten Streu

ung des Schätzers ein. Damit wird auch deutlich, daß letztere um so geringer ausfallen wird, je homogener die Schichten intern sind. Zum anderen ist deutlich, daß die Varianz des Schätzers innerhalb jeder Schicht mit einem beliebigen (für die Art der Stichprobenziehung innerhalb der Schicht angemessenen) Schätzer bestimmt werden kann. Das kann - unter der Annahme einfacher Zufallsstichproben innerhalb der Schichten - das einfache arithmetische Mittel und der bekannte Standardfehler sein. Gibt es z.B. unterschiedliche Auswahlwahrscheinlichkeiten innerhalb der Schichten, dann muß wiederum ein gewichtetes Mittel und der dafür angemessene Standardfehler verwendet werden.

Klumpung

Dadurch, daß die Auswahl von Befragungspersonen (innerhalb der Schichten) bei komplexen Stichprobendesigns oft in mehrstufigen Verfahren aus räumlich zusammenhängenden Gebieten (Klumpen, clustern) erfolgt, enthalten derartige Stichproben weniger Informationen über die Variabilität der erhobenen Merkmale in der Grundgesamtheit als einfache Zufallsstichproben. Dies gilt selbstverständlich um so mehr, je homogener die Befragtenangaben *innerhalb* der einzelnen Klumpen sind und je größer die Heterogenität *zwischen* den Klumpen ist. Würden - als extremes Beispiel - alle Befragungspersonen in einem ausgewählten Stimmbezirk die gleichen Angaben zu einer Frage machen, würden die Angaben einer einzigen Person ausreichen, um die Informationen aus diesem Stimmbezirk wiederzugeben. Wenn man eine Stichprobe mit Klumpungseffekten deshalb so behandelt, als wäre sie eine einfache Zufallsstichprobe, so ignoriert man damit diesen Genauigkeitsverlust und schließt von der geringeren Varianz der Merkmalsausprägungen in der Klumpenstichprobe fälschlicherweise auf eine entsprechend geringe Merkmalsvarianz in der Grundgesamtheit. Damit wird dann aber (wenn die Annahme der höheren internen Homogenität zutrifft) auch die daraus abgeleitete Fehlervarianz (bzw. der Standardfehler) unterschätzt.

Angenommen wir hätten aus einer Grundgesamtheit von A gleich großen Klumpen der Größe B eine einfache Zufallsstichprobe von a Klumpen gezogen und in jedem dieser Klumpen *alle Elemente ausgewählt*. Bezeichnen wir die Ausprägung des Merkmals Y für Element β ($\beta = 1,2,...,B$) in Klumpen α ($\alpha = 1,2,...,A$) mit $y_{\alpha\beta}$, so könnten wir *für jeden der Klumpen* das arithmetische Mittel eines Merkmals Y bestimmen als :

$$\bar{y}_{\alpha} = \frac{\sum_{\beta=1}^{B} y_{\alpha\beta}}{B}$$

Bei gleich großen Klumpen entspricht der Stichprobenmittelwert des Merkmals Y dem einfachen Mittelwert aus den a gezogenen *Klumpenmittelwerten*. Man kann also die einfache Zufallsstichprobe von a Klumpen als einfache Zufallsstichprobe (ohne Zurücklegen) von a Mittelwerten aus der Grundgesamtheit von A Mittelwerten auffassen. Daraus folgt, daß

$$\bar{y}_c = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{B} y_{\alpha\beta}}{n} = \frac{\sum_{\alpha=1}^{a} \bar{y}_\alpha}{a}$$

ein unverzerrter Schätzer für das arithmetische Mittel $\bar{Y}$ in der Grundgesamtheit ist. Unter den genannten Voraussetzungen ist ein Schätzer für die Fehlervarianz des arithmetischen Mittels (***Kalton*** 1983: 30) der Klumpenstichprobe $\bar{Y}_c$ gegeben als:

$$\hat{\sigma}^2_{\bar{Y}_c} = \left(1 - \frac{a}{A}\right)\frac{s_a^2}{a} \qquad \text{mit } s_a^2 = \sum_{\alpha=1}^{a} \frac{\left(\bar{y}_\alpha - \bar{y}\right)^2}{a-1}$$

An dieser Berechnungsweise ist für unsere Zwecke insbesondere bedeutsam, daß die Fehlervarianz auf der Basis der *Streuung der Mittelwerte* und *nicht* auf der Basis der Individualwerte geschätzt wird[9]. Je heterogener die Klumpenmittelwerte sind, desto größer ist somit die geschätzte Fehlervarianz. Klumpenmittelwerte sind um so heterogener, je geringer die Steuung der Merkmale innerhalb der Klumpen ist. Somit erreichen wir mit dieser Berechnungsweise eine Berücksichtigung von Klumpungseffekten.

Allerdings sind die in den obigen Ausführungen gemachten Annahmen für tatsächliche Klumpenstichproben, wie sie in der Praxis der empirischen Sozialforschung verwendet werden, unrealistisch. Normalerweise werden die Klumpen (z.B. Wahlbezirke) nicht gleich groß sein. Ebenso wird man in der Regel nicht alle Elemente der gezogenen Klumpen in die Stichprobe aufnehmen, sondern innerhalb der ausgewählten Klumpen erneut eine Auswahl treffen. Unterschiede in der Größe der Klumpen berücksichtigt man in der Regel bereits auf der Ebene ihrer Auswahl durch die Festlegung von Ziehungswahrscheinlichkei-ten, die proportional zur (geschätzten) Größe sind. Hierauf werde ich weiter unten zurückkommen. Um zu verdeutlichen, welche Konsequenzen die Einführung einer weiteren Auswahlstufe für die Schätzung von Standardfehlern hat,betrachten wir das einfachste Beispiel: Wir unterstellen, daß innerhalb der ausgewählten Wahlbezirke (mit Größe B) durch ein einfaches Zufallsauswahlverfahren jeweils b Elemente gezogen werden. Nach wie vor ist der nach Formel (3) berechnete Mittelwert aus den Klumpenmittelwerten ein unverzerrter Schätzer für den Mittelwert μ in der Grundgesamtheit. Allerdings ist der Klumpenmittelwert

$$\bar{y}_\alpha = \frac{\sum_{\beta=1}^{b} y_{\alpha\beta}}{b}$$

nun nicht mehr der ‚wahre' Mittelwert des Klumpens α, sondern ebenfalls ein durch eine

9 Die Berechnung der *Fehlervarianz* (nicht der Elementvarianz) auf der Basis der Streuung der Mittelwerte ist deshalb problemlos möglich, weil wir innerhalb der Klumpen *alle* Elemente ausgewählt haben. Somit gibt es innerhalb der Klumpen *keinen Stichprobenfehler* und wir können auch keine Fehlervarianz bestimmen.

Stichprobe gewonnener Schätzer. Aus diesem Grund müssen wir die Fehlervarianz aus dieser Schätzung bei der Bestimmung der gesamten Fehlervarianz unserer Mittelwertschätzung zusätzlich berücksichtigen. Man kann zeigen (***Kish*** 1965: 166ff, ***Kalton*** 1983: 33ff), daß sich die Fehlervarianz des Schätzers für das arithmetische Mittel $\overline{Y}_{ZwSt}$ bei diesem *zweistufigen* Design wie folgt unverzerrt schätzen läßt:

$$\hat{\sigma}^2_{\overline{Y}_{ZwSt}} = \left(1 - \frac{a}{A}\right)\frac{s_a^2}{a} + \frac{a}{A}\left(1 - \frac{b}{B}\right)\frac{s_b^2}{a\,b}$$

wobei s_a^2 die Varianz der Klumpenmittelwerte bezeichnet, wie sie bereits aus (4) bekannt ist. s_b^2 bezeichnet die durchschnittliche Merkmalsvarianz innerhalb der Klumpen und berechnet sich als:

$$s_b^2 = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b}\left(y_{\alpha\beta} - \overline{y}\,\text{subalpha}\right)^2}{a\,(b-1)}$$

Der linke Term in Gleichung (5) quantifiziert den Anteil der Fehlervarianz, der auf die Klumpung bei der ersten Stufe des Auswahlverfahrens zurückzuführen ist (‚between clusters component', ***Kish*** 1965: 166), der rechte Term bezeichnet den Anteil, der auf die zweite Stufe des Auswahlverfahrens zurückzuführen ist (‚within clusters component'). Wie man sich leicht überzeugen kann, wird der linke Term dieser Gleichung Null wenn a = A ist. Das ist dann der Fall, wenn alle Klumpen der Grundgesamtheit ausgewählt werden und somit als Schichten zu betrachten sind. In diesem Fall bleibt als Varianzschätzer der rechte Term übrig, und man kann leicht sehen, daß dieser Ausdruck dem Schätzer für proportional geschichtete Stichproben entspricht. Der rechte Term aus Gleichung (5) wird Null, wenn innerhalb der ausgewählten Klumpen alle Elemente ausgewählt werden und somit b = B ist. In diesem Fall geht Gleichung (5) in Gleichung (4) über.

Für die praktische Anwendung dieser Gleichung auf die Schätzung von Standardfehlern bei mehrstufigen Auswahlverfahren mit Klumpungseffekten auf der ersten Auswahlstufe ist jedoch eine weitere Eigenschaft der hierfür geschätzten Fehlervarianz bedeutsam: Wenn der *Auswahlsatz* der primären Stichprobeneinheiten a/A klein ist, so wird der rechte Term von Gleichung (5) klein - zumindest solange die anderen Größen in diesem Term nicht sehr groß werden. Häufig wird man deshalb in der Praxis den Beitrag der Streuung auf der zweiten Stufe und eventuellen weiteren Stufen des Auswahlverfahrens *vernachlässigen können.* Das führt zu folgendem einfachen Schätzer für die Varianz von mehrstufigen Klumpenstichproben:

$$\hat{\sigma}^2_{\overline{Y}_{ZwSt}} = \frac{s_a^2}{a}$$

Unter dieser Näherung spielt es also auch keine Rolle,mit welchem Design die Auswahl auf der zweiten Stufe erfolgt-natürlich vorausgesetzt,daß es sich um ein Zufallsauswahlverfahren handelt. Das hier nur für zwei Auswahlstufen gezeigte Vorgehen ist somit direkt auf weitere Auswahlstufen übertragbar.Auch hier gilt,daß die durch weitere Stufen begründete Variabilität oft vernachlässigt werden kann,wenn der Auswahlsatz auf der ersten Stufe klein ist. Für Stichproben nach dem ADM-Stichprobendesign umfaßt die Grundgesamtheit der Sampling Points auf der ersten Auswahlstufe 67.563 (synthetisierte) Wahlbezirke (***Arbeitsgemeinschaft ADM-Stichproben*** und ***Bureau Wendt*** 1994: 201)[10]. Der Auswahlsatz auf der ersten Stufe des Ziehungsverfahrens betrug für den Sozialwissenschaftenbus III/1996 somit a/A = 630/ 67.563 = 0,009. Insofern scheint die durch Formel (6) erreichte vereinfachte Berechnungsweise für diese Stichprobe problemlos anwendbar zu sein.

Nichtlinearität von Schätzern

Inferenzstatistische Schlüsse von Stichproben beruhen auf der Betrachtung (mit Mitteln der theoretischen Statistik oder durch Simulationen) der zu erwartenden Verteilung von Schätzern für Parameter der Grundgesamtheit bei (theoretisch) unendlich häufiger Wiederholung der Stichprobenziehung unter *gleichen Bedingungen.*Eine der konstant zu haltenden Bedingungen ist dabei der Stichprobenumfang. Wie ***Kalton*** (1984: 38ff.) jedoch anschaulich zeigt, hängt der (erwartete) Stichprobenumfang von Stichproben aus einer Grundgesamtheit von Klumpen verschiedener Größe von der Festlegung der korrekten Auswahlwahrscheinlichkeiten der einzelnen Klumpen ab. Nur wenn die Umfänge der einzelnen Klumpen am Tag der Stichprobenziehung in der Grundgesamtheit *exakt* bekannt wären, wäre auch eine fehlerfreie Bestimmung von geeigneten Auswahlwahrscheinlichkeiten für ein PPS Design möglich. Da diese Voraussetzung in der Praxis so gut wie nie (ganz sicher nicht beim ADM-Design) erfüllt sein dürfte, handelt es sich bei dem Auswahlverfahren für die (synthetisierten) Wahlbezirke um ein Verfahren, daß treffender als PPES (probability proportional to *estimated* size) zu beschreiben ist.

Diese Abhängigkeit der für jede PSU in der Grundgesamtheit festgelegten Auswahlwahrscheinlichkeit von der Schätzung ihres Umfanges macht den erwarteten Stichprobenumfang von den konkret ausgewählten Wahlbezirken abhängig. Somit muß der realisierte Stichprobenumfang von Stichproben nach dem PPES-Design als das Ergebnis eines Zufallsexperiments betrachtet werden. Als Konsequenz daraus ist z.B. der Punktschätzer für das arithmetische Mittel nicht mehr als die Division einer Zufallsvariablen (ΣX_i) durch eine Konstante (den Stichprobenumfang n) aufzufassen. Statt dessen ist auch der Nenner dieses

10 Man kann darüber debattieren, ob der angesetzte Auswahlsatz der tatsächlichen Auswahlgesamtheit Rechnung trägt, da GFM-GETAS/WBA lediglich drei ihrer Stichprobennetze für die Auswahl eingesetzt hat. Insofern wird faktisch aus einer kleineren Anzahl von Sampling Points ausgewählt. Allerdings sind die auf die am ADM-Verfahren teilnehmenden Institute aufgeteilten Netze überschneidungsfrei.

Schätzers eine Zufallsvariable, und das arithmetische Mittel wird somit durch einen Bruch aus zwei Zufallsvariablen - einen sogenannten Verhältnisschätzer - geschätzt. Weitere Zufallseinflüsse auf den Nenner des Verhältnisschätzers resultieren aus der Verwendung von Gewichtungsvariablen zum Ausgleich unterschiedlicher Auswahlwahrscheinlichkeiten auf anderen Stufen des Auswahlverfahrens bzw. zur Anpassung an Randverteilungen. Auch in die Bestimmung der Gewichtungsfaktoren fließen Schätzungen von Größen der Grundgesamtheit ein.

Die Fehlervarianz von Verhältnisschätzern ist im Gegensatz zur Fehlervarianz von geläufigen linearen Schätzern nicht problemlos auf der Basis der Merkmalsvarianz des interessierenden Merkmals in der Stichprobe schätzbar. Die statistische Theorie kennt jedoch eine Reihe von Verfahren, mit denen unter bestimmten Bedingungen gute Näherungen für die Fehlervarianz von Verhältnisschätzern gewonnen werden können. Eine verbreitete Lösung - wie sie unter anderem auch in STATA implementiert ist - ist die Verwendung von Näherungsformeln zur Linearisierung von nichtlinearen Funktionen (der erwähnte Verhältnisschätzer für das arithmetische Mittel ist eine solche nichtlineare Funktion). Für die Varianzschätzung auf der Basis von komplexen Stichprobenplänen kann man sich z.B. einer Taylor-Reihe 1. Ordnung bedienen (***Kalton*** 1983: 44f., ***Skinner*** et al. 1989: 50f). Dieses auch unter dem Namen Delta-Methode bekannte Verfahren führt unter der Voraussetzung nicht zu kleiner Stichproben und einer begrenzten Variabilität des Nennerausdrucks zu unverzerrten Schätzern für die Fehlervarianz des Verhältnisschätzers. Die Bestimmung der entsprechenden Linearisierungsfunktion setzt weiterhin voraus, daß der Schätzer, dessen Fehlervarianz bestimmt werden soll, als differenzierbare Funktion ausgedrückt werden kann (was z.B. für den Median nicht der Fall ist).

Skinner et al. (1989: 51ff) diskutieren als Alternativen zu diesem Linearisierungsansatz verschiedene Techniken, die auf der Replikation des Auswahlprozesses basieren. Zu nennen wären hier z.B. einfache Replikation, balancierte wiederholte Replikation und Jackknife-Verfahren. Diesen Verfahren ist gemeinsam, daß sie die Streuung der Schätzergebnisse zwischen den Replikationen zur Schätzung der Fehlervarianz heranziehen. Der wichtigste Vorzug dieser Verfahren ist ihre Allgemeinheit und damit die Möglichkeit, Fehlervarianzen auch für extrem komplexe (nichtlineare) Schätzer bestimmen zu können. Auf der anderen Seite müssen weitere Voraussetzungen erfüllt sein, damit diese Verfahren anwendbar sind. So ist z.B. die balancierte wiederholte Replikation für Stichproben mit vielen Schichten und idealerweise zwei PSUs pro Schicht ausgerichtet. ***Skinner et al.*** (1989: 52) zeigen auf der Basis von Simulationen, daß für die Methode der einfachen Replikation sehr viele Replikationen erforderlich sind, um ähnlich präzise Schätzer zu erlangen wie mit der Linearisierungsmethode. Welches Verfahren zur Schätzung von Standardfehlern bei komplexen Stichproben das jeweils am besten geeignete ist, wird durchaus kontrovers diskutiert.

Maximum-Likelihood-Schätzer und komplexe Stichproben

Im Prinzip sind die oben diskutierten Modifikationen des Standardverfahrens zur Schätzung von Fehlervarianzen auch auf viele Schätzer - wie z.B. Logit- oder Probitkoeffizienten - anwendbar, die mit Maximum-Likelihood-Schätzverfahren bestimmt werden. Allerdings ist dabei zu berücksichtigen, daß derartige Schätzer die statistische Unabhängigkeit der Beobachtungen voraussetzen (u.a. ***Eliason*** 1993: 7). Gerade diese Annahme ist jedoch bei Klumpenstichproben in der Regel verletzt. Aus diesem Grund kann man nicht mehr zeigen, daß die spezifizierte Likelihoodfunktion im statistischen Sinn korrekt ist. Dennoch lassen sich mit einem Standardvorgehen, das um die Berücksichtigung unterschiedlicher Auswahlwahrscheinlichkeiten durch Gewichtung erweitert ist, unverzerrte Schätzer für die gesuchten Koeffizienten gewinnen. Das Schätzverfahren bezeichnet man dann allerdings nicht mehr als Maximum-Likelihood-Schätzung, sondern als *Pseudo-Maximum-Likelihood-Schätzung.* Neben unverzerrten Punktschätzern lassen sich mit diesem Schätzverfahren auch Fehlervarianzen und Kovarianzen der geschätzten Koeffizienten gewinnen. Dabei kann ebenso wie bei anderen Schätzern das komplexe Stichprobendesign berücksichtigt werden. Aus diesem Grund sind Hypothesentests für einzelne Koeffizienten nach der gleichen Logik durchführbar wie bei einfachen Zufallsstichproben.

Der üblicherweise für Tests der Modellanpassung und häufig für vergleichende Tests von hierarchischen Modellen eingesetzte Likelihood-Verhältnis-Test ist jedoch nicht robust gegenüber der unkorrekten Likelihood-Funktion. Deshalb verbietet sich bei Pseudo-Maximum-Likelihood-Schätzern die Interpretation von Likelihood-Werten. Als Alternative zu Likelihood-Verhältnis-Tests kann man jedoch auf angepaßte Wald-Tests zurückgreifen. Die übliche Wald-Statistik $W = \boldsymbol{\beta}' \mathbf{S}^{-1} \boldsymbol{\beta}$ folgt jedoch nicht mehr einer Chi²-Verteilung, wenn die geschätzten Koeffizienten und deren Standardfehler auf einer komplexen Stichprobe basieren. ***Korn/Graubard*** (1990) können aber zeigen, daß sich die Verteilung einer Transformation dieser Statistik durch eine F-Verteilung aproximieren läßt. Hiernach folgt der Ausdruck

$$\frac{W}{d} * \frac{d - p + 1}{p}$$

näherungsweise einer F-Verteilung mit $df_1 = p$ Zählerfreiheitsgraden und $df_2 = (d - p + 1)$ Nennerfreiheitsgraden. Dabei entspricht W der oben definierten Wald-Statistik auf der Basis des für die komplexe Stichprobe geschätzten Vektors von Koeffizienten $\boldsymbol{\beta}$ und der entsprechenden Varianz-Kovarianz-Matrix $\mathbf{S}$, p der Anzahl der simultan zu testenden Koeffizienten und d der Anzahl der gezogenen primären Stichprobeneinheiten minus der Anzahl der Schichten.

4 Macht es einen Unterschied?

Nach all der statistischen Theorie blieb bislang die Frage unbeantwortet, inwieweit die Verwendung der im vorangegangenen Abschnitt umrissenen Verfahren in der empirischen Forschungspraxis tatsächlich zu anderen Ergebnissen führen als die üblicherweise verwendeten Schätzer. Schließlich wird es kaum jemanden überzeugen, die reine statistische Lehre zu vertreten, wenn sich die Ergebnisse für alle praktischen Zwecke allenfalls in den Nachkommastellen unterscheiden. Neben einem kurzen Rückblick auf das in der Einleitung verwendete Beispiel möchte ich an zwei konkreten Beispielen für Analysen mit dem Sozialwissenschaftenbus III/96 zeigen, daß es sehr wohl einen (in den Signifikanzen) deutlich spürbaren Unterschied macht, mit welchen Verfahren die Standardfehler und die daraus abgeleiteten Teststatistiken geschätzt werden.

Kommen wir zunächst auf die Schätzung des durchschnittlichen persönlichen Nettoeinkommens aus dem einleitenden Beispiel zurück. Dort hatte ich argumentiert, daß die gewichtete Berechnung des Punktschätzers zu einem unverzerrten Schätzer für das Durchschnittseinkommen in der Grundgesamtheit führt, die gewichtete Berechnung der Fehlervarianz jedoch die Einflüsse des Stichprobenplans außer acht läßt. Aus diesem Grund ist zu vermuten, daß der in Formel (2) geschätzte Standardfehler unter Berücksichtigung der tatsächlichen Stichprobenstruktur des Sozialwissenschaftenbusses anders ausfällt. Schätzt man das arithmetische Mittel deshalb mit der Prozedur *svymean* von STATA 5.0 und spezifiziert dabei jeweils eine identifizierende Variable für die Schichten (Strata) und die primären Stichprobeneinheiten (PSUs) sowie eine Gewichtungsvariable (Probability Weight, pweight), so erhält man folgendes Ergebnis:

```
Survey mean estimation

pweight:  wichtges                Number of obs    =      2660
Strata:   ost                     Number of strata =         2
PSU:      psu                     Number of PSUs   =       604
                                  Population size  =  2088.895

--------------------------------------------------------------
Mean     | Estimate   Std. Err.  [95% Conf. Interval]   Deff
---------+----------------------------------------------------
perseink |  1892.846  43.06036    1808.28   1977.413  1.944916
--------------------------------------------------------------
```

Wie nicht anders zu erwarten, ist der Punktschätzer (Estimate) dieser Schätzung mit der gewichteten Berechnung in (1) identisch. Der geschätzte Standardfehler (Std. Err.) ist jedoch *erheblich größer* als der unter einfacher Übertragung der Formeln für einfache Zufallsstichproben in (2) geschätzte Standardfehler. Dieser Sachverhalt ist auch am sogenannten Designeffekt *Deff* (***Kish*** 1965: 161) ablesbar:

$$\text{Deff} = \frac{\hat{\sigma}^{2}_{\bar{x}_{\text{kompl}}}}{\hat{\sigma}^{2}_{\bar{x}_{w}}} = \frac{43,06^{2}}{30,88^{2}} = 1,94$$

Eine anschauliche Interpretation dieses Designeffektes besteht im Rückbezug auf den Stichprobenumfang: Um mit dem komplexen Stichprobendesign eine genauso präzise Schätzung des arithmetischen Mittels zu erreichen wie mit einer einfachen Zufallsstichprobe, müßte der Stichprobenumfang der komplexen Stichprobe um den Faktor 1,94 erhöht werden.

Als zweites Beispiel für unterschiedliche Ergebnisse zwischen dem verbreiteten Vorgehen und der Berücksichtigung der Stichprobenstruktur soll ein Test auf Mittelwertunterschiede des persönlichen Nettoeinkommens zwischen Personen aus Gebieten mit unterschiedlichem Urbanisierungsgrad durchgeführt werden. Eine Möglichkeit zur Prüfung von Unterschiedshypothesen in bezug auf das Durchschnittseinkommen ist die Spezifikation eines multiplen Regressionsmodells mit Dummy-Variablen für die Kategorien der kategorialen unabhängigen Variablen ‚Urbanisierungsgrad'. Für dieses Beispiel sollen nur drei verschiedene (zusammengefaßte) BIK-Gemeindegrößenklassen zur Kennzeichnung des Urbanisierungsgrades unterschieden werden. Zunächst sind also zwei entsprechende Dummy-Variablen (D_1 und D_2) zu bilden:

Tabelle 1: Dummy-Kodierung der Gemeindegrößenklasse

BIK- Gemeindegrößenklasse		D_1	D_2	n
kleiner 50.000	‚Dorf'	1	0	436
zwischen 50.000 und 100.000	‚Mittelstadt'	0	1	731
größer 100.000	‚Großstadt'	0	0	1493

Schätzt man eine lineare Regression des persönlichen Nettoeinkommens auf die beiden Dummy-Variablen als unabhängige Variablen, so schätzt der für den Dummy D_1 geschätzte Regressionskoeffizient den Unterschied des Durchschnittseinkommens zwischen Personen aus Großstädten (der Referenzkategorie) und Personen aus ländlichen Wohnorten. Der Regressionskoeffizient für D_2 ist dementsprechend ein Schätzer für den erwarteten Einkommensunterschied zwischen Personen aus Mittelstädten und Personen aus Großstädten.

Tabelle 2: Regressionsergebnisse im Vergleich

Koeffizient[a]		SAS 6.11, *proc reg*[b] Standard-Fehler	t[d]	p > | t |	STATA 5.0, *svyreg*[c] Standard-Fehler	t[d]	p > | t |
Konstante	1978,85	42,05	47,1	0	65,46	30,2	0
D_1	-257,49	88,62	-2,9	0,004	114,21	-2,3	0,025
D_2	-149,55	69,8	-2,1	0,032	91,8	-1,6	0,104
Modellanpassung							
R^2		0,0039			0,0039		
F-Statistik (df_1, df_2)		5,16 (2, 2657)			2,84 (2, 601)		
Prob > F		0,0058			0,0595		

Anmerkungen: a) Die geschätzten Regressionskoeffizienten sind bei beiden Schätzungen identisch.
b) gewichtete Regressionsschätzung.
c) gewichtete Regressionsschätzung unter Berücksichtigung der Stichprobenstruktur.
d) Wert der Teststatistik für die Nullhypothese, daß der entsprechende Koeffizient gleich Null ist.

Quelle: Sozialwissenschftenbus III/96, gewichtete Ergebnisse

Vergleicht man die geschätzten Standardfehler dieser beiden Regressionsschätzungen, so erkennt man deutlich, daß die Berücksichtigung der Stichprobenstruktur zu erheblich größeren Standardfehlern führt. In diesem Fall hat diese Tatsache auch Konsequenzen für die Hypothesentests: Legt man eine Irrtumswahrscheinlichkeit von $\alpha = 0,05$ zugrunde, ist der Einkommensunterschied zwischen Personen aus Mittelstädten und Großstädten (Koeffizient D_2) zwar bei der gewichteten Schätzung mit SAS, nicht aber unter Berücksichtigung der Stichprobenstruktur signifikant. Das gleiche gilt für das Ergebnis des F-Tests für das Gesamtmodell. Nun kann man gegen dieses Modell einwenden, daß es in Anbetracht des extrem niedrigen R^2-Wertes ohnehin keine gute Modellierung des persönlichen Nettoeinkommens liefert. Dieser Einwand ist selbstverständlich berechtigt. Es war aber auch nicht Ziel dieser Regression, einen großen Anteil der Variabilität des Einkommens aufzuklären. Es ging ausschließlich um einen Test auf Mittelwertunterschiede. Hierfür zeigt dieses Beispiel, daß die inferenzstatistischen Schlüsse erheblich von der Art der Varianzschätzung abhängen.

Ein weiteres Beispiel soll verdeutlichen, wie sehr die gewonnenen Ergebnisse von der gewählten Art der Varianzschätzung abhängen können. Zur Untersuchung der Frage, inwieweit sich die Bewertung der Notwendigkeit von verschiedenen Merkmalen des alltäglichen Lebensstandards nach einer Reihe von vermuteten Einflußfaktoren unterscheiden,wurde für jedes der betrachteten Merkmale ein logistisches Regressionsmodell geschätzt. Abhängige

Variable dieser Schätzung ist eine dichotome Variable, die anzeigt, ob die Person das jeweilige Merkmal als notwendig für die Sicherstellung oder Erzielung eines normalen, ausreichend guten Lebensstandards in Deutschland hält. Als vermutete Einflußfaktoren wurden folgende unabhängige Variablen berücksichtigt: Eine Unterscheidung zwischen Befragten aus den neuen und den alten Bundesländern, der Wohnort des Befragten, in der oben bereits verwendeten Differenzierung nach drei verschiedenen Urbanisierungsgraden, eine Haushaltstypologie (7 Typen nach Alter und Familienstand), vier Altersklassen, das Haushalts-einkommen (Äquivalenzeinkommen), und die Schulbildung des Befragten in drei Kategorien. Dabei wurden die kategorialen Variablen durch entsprechende Dummy-Variablen kodiert. Einzelheiten hierzu und eine inhaltliche Begründung sowie Interpretation der Ergebnisse finden sich in ***Lipsmeier*** (1998).

In Tabelle 3 finden sich die Teststatistiken zur Überprüfung der Hypothese, daß die Variablen „Wohnort„, „Haushaltstyp„ und „Schulbildung„ insgesamt einen signifikanten Einfluß auf die Notwendigkeitsbewertung haben. Da diese Variablen durch mehrere Dummy-Variablen in die logistische Regression eingingen, wurden Wald-Tests verwendet. Die Tests wurden zum einen nach einer (gewichteten) Schätzung mit einer ‚normalen' Logitprozedur (*logit* und *test* von Stata 5.0) durchgeführt. Diese Ergebnisse sind in den mit **A** überschriebenen Spalten ausgewiesen. Zum anderen wurden die gleichen Tests unter Berücksichtigung des Stichprobenplans nach einer Schätzung mit der hierfür geeigneten Prozedur (*svylogit* und *svytest)* wiederholt (Spalten **B**). Bei der Interpretation der Ergebnisse ist zu beachten, daß die Teststatistiken dieser beiden Tests nicht direkt vergleichbar sind, da es sich bei der zweiten Schätzung um *angepaßte* Wald-Tests handelt. Es sollten deshalb zwischen diesen Schätzungen lediglich die Signifikanzen verglichen werden. Bereits an den dünner besetzten Spalten für die zweite Schätzung[11] ist zu erkennen, daß die Schätzung unter Berücksichtigung der Stichprobenstruktur häufig dazu führt, daß man keinen signifikanten Einfluß der betrachteten unabhängigen Variablen auf die Notwendigkeitsbewertung feststellen kann. Aus Platzgründen sind in Tabelle 3 nicht für alle der untersuchten Lebensstandardmerkmale und auch nicht für alle unabhängigen Variablen entsprechende Gegenüberstellungen dargestellt. Vergleicht man die Testergebnisse für 27 Merkmale und sechs unabhängige Variablen (***Lipsmeier*** 1998), dann zeigen sich unter Vernachlässigung des Stichprobenplans für 89 dieser insgesamt 162 Tests Ergebnisse, die auf dem 5%-Niveau signifikant sind. Berücksichtigt man dagegen den Stichprobenplan und schätzt somit konservativer, so verringert sich die Zahl der auf diesem Niveau signifikanten Testergebnisse auf 43. Derartig gravierende Unterschiede sind nicht mehr als marginal zu betrachten, sondern beeinflussen selbstverständlich die inhaltlichen Schlüsse aus den durchgeführten Datenanalysen nachhaltig.

11 Nur Teststatistiken, die mindestens auf einem Signifikanzniveau von $\alpha = 0{,}1$ signifikant sind, wurden in der Tabelle dargestellt.

Tabelle 3: Unterschiede in der Einschätzung der Notwendigkeit von ebensstandardmerkmalen

	Wohnort[a]		Haushalts
Merkmal	**A**	**B**	**A**
Mindestens alle zwei Tage eine warme Mahlzeit mit Fleisch, Fisch oder Geflügel	4,8	-	13,9*
Abgenutzte, aber noch funktionsfähige Möbel durch neue ersetzen	4,8	-	
Neue Kleidung kaufen, auch wenn die alte noch nicht abgetragen ist		2,0	15,1*
Generell mehr auf die Qualität als auf den Preis der Produkte achten können	6,5*	-	
Mindestens ein einwöchiger Urlaub weg von zu Hause pro Jahr	13,3**		
In einer guten Wohngegend leben			16,2*
Eine abgeschlossene Berufsausbildung haben	10,0**		
Alle zwei Wochen einmal abends ausgehen	8,4*	3,2**	27,6**
Ein Videorecorder	7,6*		
Ein Farbfernseher			
Mindestens einmal im Jahr ein Geschenk für Familie/Freunde kaufen können	12,7**	-	15,9*
Sich ein Hobby leisten können	10,1**	3,2**	23,7**
In der Nähe der Wohnung ein Lebensmittelladen	13,7**		
In der Nähe der Wohnung eine Bank oder Sparkasse	16,9**		
In der Nähe der Wohnung ein Arzt oder eine Ärztin	16,1**		

Anmerkungen: **A** Standard Wald-Statistiken (näherungsweise CHI² -verteilt) bei gewichteter Schätzung *hne Berücksichtigung des Stichprobenplans*

B Angepaßte Wald-Statistiken (näherungsweise F-verteilt) bei gewichteter Schätzung *mit Berücksichtigung des Stichprobenplans*

a) zwei Dummyvariablen zur Differenzierung nach 3 Urbanisierungsgraden, b) 6 Dummies für 7 Haushaltstypen, c) 2 Dummies

d) Test des Gesamtmodels mit insgesamt 15 unabhängigen (Dummy-)Variablen.

Signifikanzen: ** $\alpha \leq 0{,}01$; * $\alpha \leq 0{,}05$; ohne Stern $\alpha \leq 0{,}1$; nicht signifikante Koeffizienten sind nicht ausgewiesen; [ ] Modellfit auf

5% Niveau nicht signifikant; - Koeffizient nicht ausgewiesen da der Modellfit insgesamt nicht signifikant ist

Quelle: Sozialwissenschaftenbus III/96, gewichtete Ergebnisse

5 Fazit und geeignete Software

Wenn durch die Argumentation und die Beispiele dieses Beitrags deutlich geworden ist, daß das Auswahlverfahren von in der sozialwissenschaftlichen Forschung verwendeten Umfragedaten einen erheblichen Einfluß auf die Fehlervarianzen von verschiedenen Schätzern haben kann, so stellt sich die Frage nach den Konsequenzen. Ein Ziel dieses Beitrages war es aufzuzeigen, daß es mit moderner Software, wie z.B. STATA 5.0 relativ problemlos möglich ist, die Stichprobenstruktur bei den eigenen Datenanalysen zu berücksichtigen. Allerdings setzt das - neben der Verfügbarkeit geeigneter Software - voraus, daß die Forscher und Forscherinnen in ihren Datensätzen über die notwendigen Informationen verfügen. Dazu gehört neben einer ausführlichen Dokumentation des Auswahlverfahrens auch die Bereitstellung identifizierender Variablen für Schichten und primäre Stichprobeneinheiten sowie entsprechender Gewichtungsfaktoren (wenn solche erforderlich und/oder gewünscht sind). Daß diese Voraussetzung leider nicht immer erfüllt ist, zeigt das Beispiel des europäischen Haushaltspanels. Zwar sind die entsprechenden Daten der ersten drei Wellen inzwischen für die Wissenschaft zugänglich, das Statistische Bundesamt ist jedoch aus Datenschutzgründen bei der Weitergabe der faktisch anonymisierten Daten nicht bereit, eine identifizierende Variable für die primären Stichprobeneinheiten mitzuliefern.

Stehen die notwendigen Informationen zur Verfügung, so ist zu klären, welche Software für die angestrebten Datenanalysen geeignet ist. Meines Wissens verfügen bislang weder SPSS noch SAS in der jeweiligen Standardversion über geeignete Prozeduren zur Berücksichtigung von komplexen Stichprobenplänen. Wie oben bereits erwähnt, sind mittlerweile jedoch einschlägige Ergänzungen zu diesen Programmen erhältlich. STATA bietet seit Version 5.0 eine Reihe von Analysemöglichkeiten, die zudem laufend erweitert werden. Informationen über die in der soeben erschienenen Version 6.0 enthaltenen Prozeduren lassen sich bequem im Internet (http://www.stata.com) abrufen. Obwohl dort bislang bereits für eine Vielzahl von Schätzprozeduren Varianten für komplexe Stichproben implementiert sind, bleiben natürlich Analyseprobleme, für die (noch) keine entsprechende Variante verfügbar ist. Das betrifft bislang insbesondere Prozeduren zur Analyse von Längsschnittfragestellungen. Es bleibt zu hoffen, daß die zunehmende Integration von entsprechenden Schätzverfahren in statistische Standardsoftware dazu beiträgt, daß diese Verfahren in Zukunft eine größere Aufmerksamkeit durch die wissenschaftliche Öffentlichkeit erfahren.

Literatur

Arbeitsgemeinschaft ADM-Stichproben und ***Bureau Wendt*** (1994): Das ADM-Stichproben- System (Stand 1993). In: ***Gabler, S./Hoffmeyer-Zlotnik, J.H.P./Krebs, D.*** (Hrsg.): Gewichtung in der Umfragepraxis. Opladen: Westdeutscher Verlag

Diekmann, Andreas (1995): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. Reinbek bei Hamburg: Rowohlt Taschenbuch

Eliason, Scott R. (1993): Maximum Likelihood Estimation: Logic and Practice. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-096. Newbury Park, CA: Sage

GFM-Getas/WBA (Hrsg.) (1997): Sozialwissenschaftenbus III/96 - Methodendokumentation zur technischen Organisation und Durchführung. Hamburg: GFM-GETAS/WBA

Kalton, Graham (1983): Introduction to Survey Sampling. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-035. Newbury Park et al.: Sage

Kirschner, Hans-Peter (1984a): ALLBUS 1980: Stichprobenplan und Gewichtung. In: ***Mayer, Karl Ulrich*** /***Schmidt, Peter*** (Hrsg.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Beiträge zu methodischen Problemen des ALLBUS 1980. Frankfurt am Main/ New York: Campus

Kirschner, Hans-Peter (1984b): Zu Stichprobenfehlerberechnungen im Rahmen des ADM-Stichprobenplans. In: ZUMA Nachrichten 15, S. 40-71.

Kish, Leslie (1965): Survey sampling. New York et al.: Wiley

Korn, Edward L./Graubard, Barry I. (1990): Simultaneous Testing of Regression Coefficients With Complex Survey Data: Use of Bonferoni t Statistics. In: The American Statistician, Vol. 44, No. 4, S. 270 - 276.

Lipsmeier, Gero (1998): Wie homogen sind die Vorstellungen darüber, was in Deutschland zum Lebensstandard gehört? Zur Möglichkeit einer ‚demokratischen' Bestimmung des notwendigen Lebensstandards mit Umfragedaten. Arbeitspapier Nr. 5 des Projektes „Indikatoren für die Wohlfahrtsposition von Haushalten - Deprivationsbasierte Armutsmaße„. Bielefeld: Universität Bielefeld, Fakultät für Soziologie

SAS Institue Inc. (1990): SAS Procedures Guide, Version 6, Third Edition. Cray, NC: SAS Institute Inc.

Schnell, Rainer/Hill, Paul B./Esser, Elke (1992): Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg (3. Auflage)

Skinner, C.J./Holt, D./Smith, T.M.F. (1989): Analysis of Complex Surveys. Chichester et al.: John Wiley & Sons

Statistisches Bundesamt (Hrsg.) (1998): Bevölkerung und Erwerbstätigkeit. Fachserie 1, Reihe 1. Gebiet und Bevölkerung 1. bis 4. Vierteljahr 1996. Wiesbaden

Stenger, Horst (1986): Stichproben. Heidelberg/Wien: Physica Verlag